Gridded population data for the UK – redistribution models and applications

David Martin[1]
[1]School of Geography, University of Southampton, Southampton, SO17 1BJ, UK e-mail: D.J.Martin@soton.ac.uk

Abstract

This paper will review the methodology developed for the production of grid-based population estimates for the UK - a country where the national statistical organization has only once before experimented with this form of output, in 1971. It has therefore been necessary for researchers to use ancillary information from census area centroids and boundaries in order to redistribute published census totals onto regular grid geographies. The paper will consider the way in which gridded models have been produced from published aggregate census data examine the applications in which these data models have been used, demonstrating several areas in which they offer unique analysis opportunities which cannot be addressed using irregular area-based demographic data. This work is placed in the context of current plans by the Office for National Statistics to create a new grid square product using unpublished small area population estimates.

Keywords: grid map, population, data

1. Introduction.

Despite the broad utility of grid-based models of population distribution, there has been limited availability of such representations in the UK. A national 1-km grid data product was produced as an output of the 1971 census and provided the basis for a census atlas of Britain (CRU/OPCS/GROS, 1980). The 1971 gridded data were based on direct aggregation of geographically referenced census returns to grid squares by the statistical organizations. However, except in Northern Ireland, where gridded outputs have been produced from each successive census, (http://cdu.mimas.ac.uk//2001/ni/grid/) this exercise was not repeated and researchers have thus not had access to any definitive grid-based mapping product from subsequent censuses, despite a generally increasing resolution of geographical referencing in population data products. In response to this situation, researchers have developed methods for the reallocation of population counts from published sources onto regular grids and these have seen use in a wide range of research applications. As a result of the European Grid Map initiative, the Office for National Statistics (ONS) is now planning to produce a grid square data product directly from otherwise unpublished small area population estimates.

The following three sections of this paper explain the methods by which published census data have been redistributed into regular grid geographies and review the

applications to which these data have been put. The fourth section considers new data availability and the relationship between these models and current ONS plans.

2. Small area census population data availability in the UK

In recent UK censuses, area-based aggregate data have been published for a range of geographical units, the smallest being known as enumeration districts (1971-91, typical population size 400) and output areas (2001, typical population size 300). There are some differences in detailed census methodology and outputs between the countries of the UK, which are not covered in detail here, but it is worthy of note that Scotland adopted output areas earlier and they have tended to be somewhat smaller than elsewhere. Over the period from 1971, aggregate population statistics have been accompanied by population-weighted centroid locations. In 1971-91 these were allocated manually by staff at the statistical organizations, while in 2001 they were computed directly from the GIS data used in the design of the census output areas, a separate process explained in more detail in Martin (2002). Digital boundaries at the smallest levels of census geography were not produced nationally in 1971 or 1981. In 1991 digital boundaries were created after the census by digitizing paper maps and in 2001 they were an integral part of the GIS-based census design. There is extensive use of the postcode as a geographical reference in the UK, with the smallest level of the postcode system typically representing 15 households. Postcodes are owned and developed by Royal Mail, the national postal service provider, to aid the delivery of mail and until 2001 there was no direct correspondence between the postcode and census geography, except in Scotland. For the first time in 2001, census output areas were wherever possible built from whole unit postcodes. Directories have been produced which identify the census zones into which each postcode falls, although individual postcodes are often split. In the most recent versions of the National Statistics Postcode Directory (www.statistics.gov.uk/geography/nspd.asp), each postcode record includes a 1m resolution grid reference, address count and allocation to multiple census and administrative geographies. It is thus possible to associate (e.g.) census output area data with a list of postcodes, each of which has a location and count of addresses. In the absence of any grid-square data product from the national statistical organizations, users requiring grid square data outside Northern Ireland in the period following the 1971 census have had to generate their own models by reallocating population data from the published census geographies into a regular grid, using some combination of these ancillary datasets.

3. Grid-based population modeling from centroid data.

A grid-based population surface modelling method was introduced by Martin (1989) and Bracken and Martin (1989), designed to address the demand for grid-based population models in situations where these needed to be produced from aggregate published outputs for irregular small areas. This method was utilised to produce a series of national population surfaces based on 1981 and 1991 census data (Bracken and Martin, 1995) and subsequently to produce a new interface and extension of the dataset to Northern Ireland (Martin et al., 2000) known as Surpop. Martin (1996) sets out the key principles underlying the approach, which is based on the redistribution of population counts from the population-weighted centroid locations, into the cells of a

regular grid using adaptive kernel estimation. These models produce a very different spatial representation to that resulting from conventional area-based mapping, most importantly reconstructing the geography of settled and unpopulated areas.

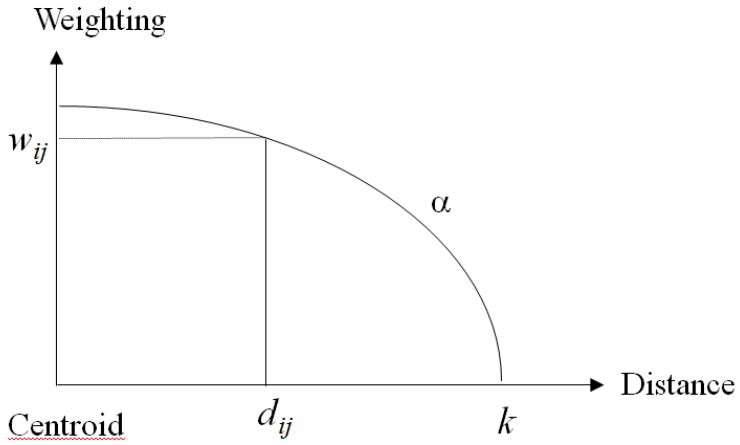
Various alternative techniques for generating surface representations of socioeconomic data have been driven by similar considerations, including Tobler (1979); Goodchild et al., (1993); Langford and Unwin (1994), Langford (2007) and Thurstain-Goodwin and Unwin (2000). Kyriakidis (2004) proposes a geostatistical framework for area-to-point interpolation and would prefer not to associate populations with centroids locations, yet this is the very data type which contemporary administrative data systems are tending to produce. A range of areal interpolation methods are available for the translation of data from one set of areal units to another incompatible set, and translation from irregular census zones to regular grid cells may be considered as a special case of such interpolation. However, the approaches identified here all differ from more the general areal interpolation techniques in that they variously embody features unique to the estimation of a spatially continuous population distribution, albeit approximated by the regular spatial grid.

The basic model proposed by Martin (1989) assumes the existence of population count data for irregularly sized and shaped zones, the boundaries of which are not always known, but which are represented by population-weighted centroid locations. Each centroid is treated as a local summary point for the more detailed, but unknown, actual population distribution. Using the UK data described above, the population of a census output area could be represented by its population-weighted centroid or, more recently, a set of associated unit postcode locations with population counts. The counts associated with these locations are redistributed into the cells of a regular spatial grid. The algorithm proceeds by focusing on each centroid in turn. Mean inter-centroid distance is determined within an initial user-defined kernel width. The kernel width is then adapted to equal this distance and weights assigned to local cells according to a distance decay function. In the construction of the Surpop models, w_{ij} the weighting of cell i with respect to centroid j, is determined by:

$$w_{ij} = \left(\frac{k^2 - d_{ij}^2}{k^2 + d_{ij}^2} \right)^{\alpha}$$

where k is the initial kernel width and d_{ij} is the distance between the centre of cell i and centroid j, as illustrated in Figure 1. The exponent α offers control over the shape of the distance decay function within the extent of the spatial kernel, although alternative distance decay functions could be readily specified. The Surpop models were based on enumeration district data and national models were produced for a grid of 200m x 200m using a simple implementation of the decay function with $\alpha = 1$. Martin (1996) introduced the use of a pre-processed mask layer in the form of a rasterized map at the same resolution as the intended output grid. This can be used to restrict the redistribution process such that weights associated with all cells which do not match the centroid currently being processed are set to zero, enforcing population volume preservation within the defined areal unit associated with each centroid (within the limitations of the grid resolution). The mask layer can also be used to protect invalid areas such as lakes or the sea from receiving population allocations.

Figure 1: General form of adaptive distance weighting kernel

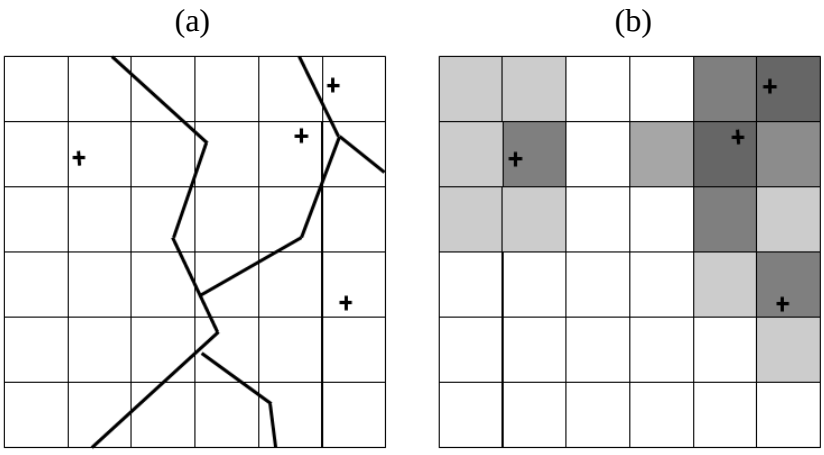


The total population recorded at each centroid is redistributed in proportion to the weights assigned to cells within the kernel, subject to any constraints from the mask layer. The total population received by cell *i* is thus the sum of its weighted population from all centroids:

$$\hat{P}_i = \sum_{j=1}^{N} P_j w_{ij}$$

where N is the total number of centroids and P_j is the population at centroid j. Some cells will receive population from many overlappimg kernels, while cells in unpopulated regions will receive none. The relationship between the input centroids and the resulting gridded population estimates is illustrated in Figure 2. In this example, only one centroid is shown within each zone but the same principle applies where a zone is represented by multiple centroids, such as a census output area with several associated postcode locations. Importantly, edge effects are avoided by processing a spatial margin of at least one kernel width greater than the output region, allowing population from centroids just beyond the region to be included, and for some of the population of centroids within *k* distance of the edge to be lost.

Figure 2: General relationship between (a) zone boundaries and population-weighted centroids, and (b) population-weighted centroids and gridded population estimates.



It is clearly possible to apply this model to a range of different input data, with the key assumption being that it is appropriate to treat the population-weighted centroid locations as high-information points about the form of the local population distribution. It is thus possible to associate a population count published for a single zone with multiple centroids, given an appropriate representation of the relationship between the points and the zone. Martin (2006) presents a series of intercensal population grids by using postcode locations as centroids and address counts as weights. There are typically 11 residential postcodes within each census output area, providing 11 summary points of local population distribution rather than the one output area centroid. Regardless of the data source, this modeling approach using a spatially adaptive kernel allows for the local redistribution of population counts around points of high information into one or more cells of the gridded model.

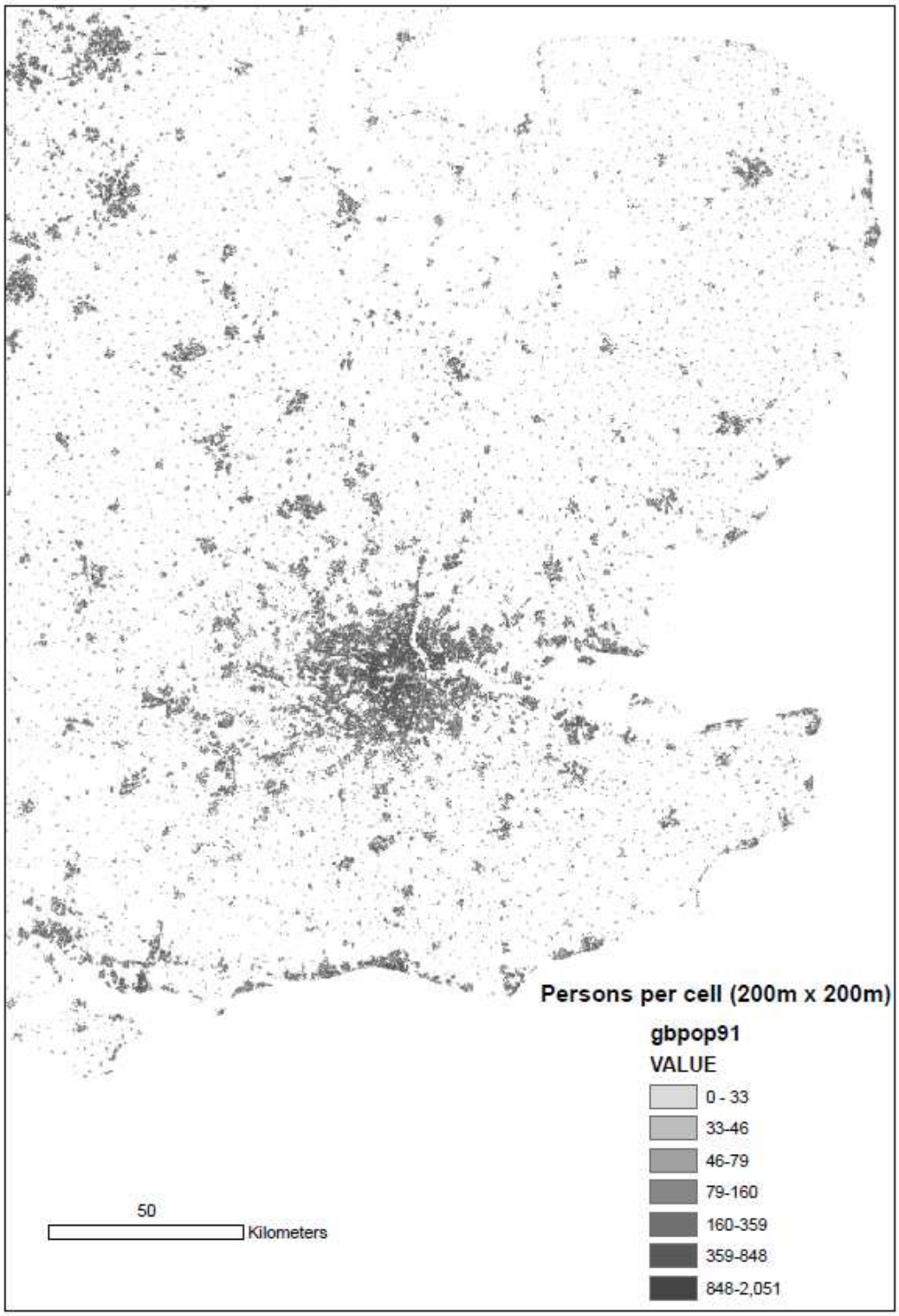
4. Applications and uses.

The national Surpop models of Bracken and Martin (1995) were produced using this method based on a 200m x 200m grid, aligned with the British National Grid. An example is illustrated in Figure 3, which covers the south east of England and clearly reveals the detailed settlement pattern. The population count for each census enumeration district was associated with a single population-weighted centroid and redistributed onto the grid as described in the preceding section. Some researchers have used these models directly, whereas others have re-applied the methodology to their own data.

Tate (2000) summarizes the general advantage of surface approaches as providing a more realistic model of settlement pattern, reflecting density changes that are hard to represent using area-based representations. Numerous researchers have used the specific method described here. These can generally be characterized by applications in which the spatial distribution of the population, independent of the zonal geography used for data publication, has an important impact on the analysis. Examples include investigation of social class inequalities in risk factors associated with flooding (Fielding, 2007); bridging the gap between disease transmission models and urbanization (Zhang and Atkinson, 2008); studies of environmental equity and risk assessment (Brainard et al., 2002; Mennis, 2003); population exposure to transport of hazardous waste (Lovett et al., 1997; Brainard et al., 1996); more general transportation cost modelling (Brainard et al., 1997; Martin et al., 2002), and as the basis for cellular automata modelling of urban expansion (Wu and Martin, 2002). Mesev et al. (1995) adopt a hybrid approach in which the census-based gridded population estimates are used in order to enhance the classification of urban areas from remotely sensed data. Martin (2006) uses gridded models from successive censuses to compare the changing geography of population where direct comparison between zonal data is not possible due to changing zonal boundaries.

In all these application examples, the comparability provided by the gridded geography is a key benefit. This takes at least two forms: (i) comparability between data sources and (ii) comparability over time. Most of the applications listed here benefit from the abstract nature of the gridded structure which allows the output of multiple data modeling exercises to be combined using a common spatial framework.

Figure 3: Gridded population model from 1991 census, southeast England, 200m x 200m grid. Source: Office for Population Censuses and Surveys, 1991 Census: Small Area Statistics (England and Wales). ESRC/JISC Census Programme, Census Dissemination Unit, Mimas (University of Manchester)



This is particularly the case with models of environmental processes such as flooding or hazard where the outputs of the process model are effectively continuous over space and represented for convenience by a regular grid. Similarly, applications

involving remotely sensed data are most readily resampled onto a regular gridded geography. The issue here is not only one of analytical convenience but also that a gridded population model at an appropriate spatial resolution is better able to convey the geography of settlement than a space-filling zonal representation. Under most settlement patterns many cells of the gridded model will have zero population, truly representing the discontinuous nature of the population distribution. The second benefit noted here is that of comparability over time. The UK in particular is subject to continual revision of census and administrative boundaries, primarily due to policies related to electoral representation. It is therefore not possible to compare the results of any of the four most recent censuses through a consistent small area geography. By contrast, the gridded representations may be directly overlaid and reveal not only the numerical change in population counts but also the changing spatial extents of populated areas.

4. New data products: small area estimates and the grid map

In addition to census datasets, ONS produce annual mid-year population estimates (MYEs). For many years, these estimates have been calculated only for large geographical areas such as local authority districts. Bates (2004) evaluates the use of administrative data sources such as health service patient registers, child benefit and older persons databases in order to produce annual estimates for much smaller geographical units, via a ratio change methodology described in Bates (2006) and ONS (2008). These data have now been published as 'experimental statistics' for the years 2001-07 for a wards and for lower and middle layer super output areas (for a fuller explanation of contemporary UK statistical geographies, see http://www.statistics.gov.uk/geography/beginners_guide.asp). The hierarchy of output areas and super output areas currently forms the basis for statistical data publication and has introduced a degree of stability for users of these datasets relative to previous continual revision of ward boundaries. The methodology underlying these small area estimates involves handling administrative data which are received at the postcode level and hence the entire system is underpinned by a set of unpublished postcode-level population estimates. These thus represent a modeled version of the postcode estimates described in section 3 above, with the very important improvements that they are annually refreshed at the postcode level using a variety of administrative sources, rather than the Martin (2006) estimates which are simply pro-rata allocations of higher-level population counts onto postcodes using numbers of addresses.

As part of the European Grid Map project, ONS are developing a 1km population grid map based on the unpublished postcode-level population estimates described above. At present, this process has only been run for England and Wales although it is hoped soon to be able to include Scotland and Northern Ireland. The initial grid is created with the GISCO tool in ArcGIS using the ETRS89 reference system. The grid map is then intersected and joined with the postcode points in order to aggregate the point-based population counts to the grid cells, resulting in the grid map. In this approach, the entire population associated with each centroid is assigned to the grid square which contains that location and there is thus no spatial smoothing of the population count in the way that was inherent to the model presented in section 3. Allocation of population counts from high resolution postcodes to 1km x 1km grid squares is

unlikely to result in large misallocation of populations between cells, although there will be situations in which a postcode centroid close to a cell boundary causes population to be allocated into the incorrect cell. This tendency would be moderated by the use of kernel estimation and masking or indeed the construction of the entire model at a higher spatial resolution prior to aggregation to 1km x 1km the grid. The GISCO grid will also be incompatible with cells of an equivalent spatial resolution based on the British National Grid, which is differently projected, thus it will not be possible to use the same cell counts in the European Grid Map and most UK-based GIS applications.

5. Conclusion.

This paper has reviewed the recent history of grid-based population mapping in the UK, explaining the absence of official grid maps with the notable exception of the 1971 census. In the interim period, researchers have developed a range of methods for the estimation of gridded population data from published census counts and a wide range of applications have been developed. Increased spatial resolution of standard products, the ready availability of GIS processing power and new data sources within the national statistical organization have now made possible the development of an experimental gridded population map. This is an exciting new development and presents a number of research opportunities and challenges. In particular, it may be possible to take advantage of the spatial smoothing features of the modeling approaches in order to enhance the new output. Further, it may be possible to use the new national-level model to calibrate modeled population distributions so that they provide estimates of the more spatially detailed pattern while preserving 'official' counts at larger scales. These gridded models potentially offer new advantages of data comparability between sources and over time, particularly for international, Europe-wide analyses. For UK GIS users, there will however be challenges in moving between the British National Grid which has previously been the basis for all published population data referencing and the harmonized European Grid Map framework. There will also be great interest in the potential use of the standard grid as a publication framework for future census and other socioeconomic data.

Acknowledgements

The author gratefully acknowledges the assistance of staff of the Office for National Statistics (ONS), particularly Andy Tait, Ross Carter and Andy Bates, and the support of Economic and Social Research Council (ESRC) Award RES-348-25-0006. All comments and views expressed here are those of the author and not necessarily of ONS or ESRC.

References

Cooper M. C., Milligan G. W. (1988) The effect of measurement error on determining the number of clusters in cluster analysis, in: *Data, Expert Knowledge and Decision*, Gaul, W. & Shader, M. (Eds.), Springer, 319-328

Bates, A. (2004) Small area population estimates project: data quality of administrative datasets *Population Trends* 116, 11-17

Bates, A. (2006) Methodology used for producing ONS's small area population estimates *Population Trends* 125, 30-36

Bracken, I. and Martin, D. (1989) The generation of spatial population distributions from census centroid data, *Environment and Planning A,* 21, 537-543

Bracken, I. and Martin, D. (1995) Linkage of the 1981 and 1991 Censuses using surface modelling concepts, *Environment and Planning A,* 27, 379-90

Brainard, J. S., Lovett, A. A. and Bateman, I. J. (1997) Using isochrone surfaces in travel-cost models, *Journal of Transport Geography,* 5 (2), 117-126

Brainard, J., Lovett, A. and Parfitt, J. (1996) Assessing hazardous-waste transport risks using a GIS, *International Journal of Geographical Information Systems,* 10 (7), 831-849

Brainard, J. S., Jones, A. P., Bateman, I. J., Lovett, A. A. and Fallon, P. J. (2002) Modelling environmental equity: access to air quality in Birmingham, England, *Environmental and Planning A,* 34(4), 695-716

CRU/OPCS/GROS (1980) People in Britain: A Census Atlas, HMSO, London

Fielding, J. (2007) Environmental injustice or just the lie of the land: an investigation of the socio-economic class of those at risk from flooding in England and Wales, *Sociological Research Online,* 12 (4)

Goodchild, M. F., Anselin, L., Deichmann, U. (1993) A framework for the areal interpolation of socioeconomic data, *Environment and Planning A,* 25 (3), 383–397

Kyriakidis, P. C. (2004) A geostatistical framework for area-to-point spatial interpolation, *Geographical Analysis,* 36 (3), 259–289

Langford, M. (2007) Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps, *Computers, Environment and Urban Systems,* 31 (1), 19-32

Langford, M., and Unwin, D. (1994) Generating and mapping population density surfaces within a geographical information system, *Cartographic Journal,* 31, 21-26

Lovett A. A., Parfitt, J. P., Brainard J. S. (1997) Using GIS in risk analysis: A case study of hazardous waste transport, *Risk Analysis,* 17 (5), 625-633

Martin, D. (1989) Mapping population data from zone centroid locations, *Transactions of the Institute of British Geographers,* NS, 14, 90-97

Martin, D. (1996) An assessment of surface and zonal models of population, *International Journal of Geographical Information Systems,* 10, 973-989

Martin, D., Tate, N. J. and Langford, M. (2000) Refining population surface models: experiments with Northern Ireland Census, Data *Transactions in GIS,* 4, 343-360

Martin, D. (2002) Geography for the 2001 Census in England and Wales, *Population Trends,* 108, 7-15
[http://www.statistics.gov.uk/downloads/theme_population/PT108.pdf]

Martin, D. (2006) Grid models of population: temporal comparison by fixing the geography. Paper presented at ESRC Research Methods Festival, University of Oxford, 17-20 July
[http://www.ccsr.ac.uk/methods/festival/programme/masc/documents/martin.ppt]

Martin, D., Wrigley, H., Barnett, S. and Roderick, P. (2002) Increasing the sophistication of access measurement in a rural healthcare study, *Health and Place,* 8(1), 3-13

Mennis, J. (2003) Generating surface models of population using dasymetric mapping, *Professional Geographer*, 55 (1), 31-42

Mesev, V., Longley, P., Batty, M. and Xie, Y. (1995) Morphology from imagery - detecting and measuring the density of urban land-use, *Environment and Planning A*, 27 (5), 759-780

ONS (2008) Methodology Note on production of Super Output Area Population Estimates, ONS [http://www.statistics.gov.uk/about/methodology_by_theme/sape/downloads/Methodology_note_SOAs.pdf]

Tate, N. J. (2000) Surfaces for GIScience Transactions in GIS 4 (4), 303-305

Thurstain-Goodwin, M. and Unwin, D. (2000) Defining and delimiting the central areas of towns for statistical monitoring using continuous surface representations, *Transactions in GIS*, 4 (4), 305-317

Tobler, W. R. (1979) Smooth pycnophylactic interpolation for geographical regions, *Journal of the American Statistical Association*, 74, 519–36

Wu, F. and Martin, D. (2002) Urban expansion simulation of Southeast England using population surface modelling and cellular automata, *Environment and Planning A*, 34, 1855-1876

Zhang P, Atkinson P. M. (2008) Modelling the effect of urbanization on the transmission of an infectious disease, *Mathematical Biosciences*, 211 (1), 166-185